Accurate predictions on small data with a tabular foundation model

Year: 2025 | Citations: 406 | Authors: Noah Hollmann, Samuel G. Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer

Abstract

Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science1,2. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although deep learning has revolutionized learning from raw data and led to numerous high-profile success stories3, 4–5, gradient-boosted decision trees6, 7, 8–9 have dominated tabular data for the past 20 years. Here we present the Tabular Prior-data Fitted Network (TabPFN), a tabular foundation model that outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. As a generative transformer-based foundation model, this model also allows fine-tuning, data generation, density estimation and learning reusable embeddings. TabPFN is a learning algorithm that is itself learned across millions of synthetic datasets, demonstrating the power of this approach for algorithm development. By improving modelling abilities across diverse fields, TabPFN has the potential to accelerate scientific discovery and enhance important decision-making in various domains. Tabular Prior-data Fitted Network, a tabular foundation model, provides accurate predictions on small data and outperforms all previous methods on datasets with up to 10,000 samples by a wide margin.